DO ESSAY AND MULTIPLE-CHOICE QUESTIONS MEASURE THE SAME THING?

by

Stephen Hickson and W. Robert Reed*

Abstract

Our study empirically investigates the relationship between essay and multiple-choice (MC) questions using a unique data set compiled from several years of university introductory economics classes. We conclude that essay and multiple-choice questions do not measure the same thing. Our main contribution is that we show that essay questions contain independent information that is related to student learning. Specifically, we find that the component of essay scores that cannot be explained by MC responses is positively and significantly related to (i) performance on a subsequent exam in the same economics course, and (ii) academic performance in other courses. A further contribution of our study is that we demonstrate that empirical approaches that rely on factor analyses or Walstad-Becker (1994)-type regressions are unreliable in the following sense: It is possible for these empirical procedures to lead to the conclusion that essay and MC questions measure the same thing, even when the underlying data contain strong, contrary evidence.

JEL Categories: A22

Keywords: Principles of Economics Assessment, Multiple Choice, Constructed Response, Free Response, Essay.

Third draft, Do not quote
March 29, 2009

*Hickson and Reed are, respectively, Teaching Fellow and Professor of Economics at the University of Canterbury. Reed is the contact author and his contact details are: Department of Economics and Finance, University of Canterbury, Private Bag 4800, Christchurch 8042, New Zealand; Email: bobreednz@yahoo.com; Phone: +64 3 364 2846. We want to acknowledge Peter Kennedy for copious comments on earlier versions of this paper. His sharp criticisms resulted in substantial improvements in our analysis.

> "In sum, the evidence presented offers little support for the stereotype of multiple-choice and free-response formats as measuring substantially different constructs."
>
> \- Bennett, Rock, and Wang (1991)

> "Whatever is being measured by the constructed-response section is measured better by the multiple-choice section…We have never found any test that is composed of an objectively and subjectively scored section for which this is not true."
>
> \- Wainer and Thissen (1993)

> "The findings from this analysis of AP exams in micro and macro principles of economics are consistent with previous studies that found no differences, or only slight differences, in what the two types of tests and questions [multiple-choice and essay] measure."
>
> \- Walstad and Becker (1994)

I. INTRODUCTION

University principles of economics courses often have enrollments of several hundred students or more. Instructors of these courses face a potential tradeoff when designing tests: On the one hand, essay questions are thought to assess important learning outcomes that are not well-addressed by multiple-choice (MC) questions. On the other hand, essay questions are much more costly to grade. In addition, the marking of essay questions is less reliable due to the subjective nature of the questions.

Ideally, one would weigh the respective benefits and costs of essay and MC questions to decide the optimal mix of each to employ. However, this is a difficult task, especially given the subjective nature of "benefits."[1] Perhaps because of this, much attention has focused on the question, "Do essay and MC questions measure the same

[1] The only study that we are aware of that attempts such an approach is Kennedy and Walstad (1997). They frame the decision to use essay questions as a tradeoff between reduced "misclassifications" and higher marking costs. "Misclassifications" are defined as estimated differences in the grade distribution (beyond natural sampling variation) that would arise on the AP micro- and macroeconomics exams from switching to an all-MC format. Unfortunately, in order to categorize these as "misclassifications," KW must assume that the mix of essay and MC questions on the AP tests is optimal. If the mix is not optimal, then it doesn't follow that the grade distribution under an all-MC format is worse than under the mixed format. This highlights the practical difficulties of implementing the "benefits versus costs" approach.

thing?" If this question could be answered in the affirmative, one could eliminate essay questions. In fact, a number of influential studies claim to demonstrate that "constructed response" and MC questions measure the same thing.[2] The implications of this have been well-understood:

> The educational measurement literature suggests that multiple-choice questions measure essentially the same thing as do constructed-response questions. Given the higher reliability and lower cost of a multiple-choice test, a good case can be made for omitting constructed-response questions from a test containing both multiple-choice and constructed-response questions because they contribute little or no new information about student achievement (Kennedy and Walstad, 1997, page 359).

Previous research has taken different approaches to this question. Bennett, Rock, and Wang (1991) and Thissen, Wainer, and Wang (1994) employ factor analysis. Walstad and Becker (1994) regress AP composite scores on MC scores. Kennedy and Walstad (1997) simulate grade distributions using different test formats. Each of these has its own notion of what it means to "measure the same thing," and none attempts to reconcile their approach to those of others.

Our study proposes its own approach to this question. We begin by asking whether essay scores are "predictable" from MC scores. If a student's performance on the essay component of a test can be perfectly, or near-perfectly, predicted by their performance on the MC component, we could easily conclude that the two components "measure the same thing."

But suppose essay scores are only imperfectly predictable from MC scores, so that the regression of essay scores on MC scores left a substantial residual. What could

[2] "Constructed response" questions, also known as "free response," are not synonymous with essay questions. The former includes any question-type where the respondent is not provided with a set of possible answers (e.g., fill-in-the-blank questions). Therefore, one should be wary of extrapolating "constructed response" findings to conclusions about the relationship between essay and MC questions. We will have a little bit to say about this further below.

we conclude from that? The innovation of our study is that we are able to demonstrate that this residual is empirically linked to student achievement. Since the residual represents the component of essay scores that cannot be explained by MC scores, and since it is significantly correlated with learning outcomes, we infer that essay questions contain "new information about student achievement" and therefore do not measure the same thing as MC questions. We then proceed to relate our results to previous research and conclude with recommendations for future research.

II. DATA

Our analysis uses data compiled over a five-year period (2003-2007) from approximately 8400 students in two different courses at the University of Canterbury in New Zealand. Introductory Microeconomics and Introductory Macroeconomics are semester-long courses typically taken by business students in their first year of study. Both courses administer a mid-semester "term test" and an end-of-semester final exam.

Both term-tests and final exams consist of an essay and a MC component. While the weights given to these components are different for the term test and the final exam, and change somewhat over the years, the structure of these components has remained constant. For both courses, the term test is 90 minutes long and consists of 25 MC and two essay questions. The final exam is longer at 180 minutes, and consists of 30 MC and three essay questions. There was little change in the coverage of the respective assessments over the years with one exception: In 2007, the final exam gave more coverage to material in the first half of the course. Inasmuch as possible, quality control across assessments was maintained by the fact that the same two instructors taught the classes, and wrote and graded the assessments across the whole time period.

All together, the data set includes assessments from ten separate offerings of Introductory Microeconomics and eight of Introductory Macroeconomics, for a total of 36 assessments (18 term tests plus 18 final exams). When we eliminate incomplete records and students for whom one of the assessments is missing, we are left with 16,710 observations.[3] By way of comparison, Walstad and Becker (1994) have a total of 8,842 observations. Most studies have far fewer.[4]

There are two features which make our data set unique. First, we have repeated observations on the same student for a given course. This allows us to test whether essay scores on the term test provide "new information" that can be used to predict student achievement on the final exam. Second, we have information about the student's achievement in other courses. This allows us to test whether essay scores in an economics course provide "new information" about student achievement outside the class.

The two key variables in our study are student scores on the essay and MC components of their term tests/final exams. These are calculated as percentages out of total possible scores. Panel A of FIGURE 1 reports a histogram and statistical summary for the full sample of essay scores. The average score is 52.53, and there is evidence of clumping as a result of the way in which the percentage scores are calculated. The lower panel of FIGURE 1 provides a similar report for the MC scores in our study. These are characterized by a higher mean (68.38) and smaller spread.

[3] The main reasons for deleting observations were the following: (i) A student received an aegrotat pass. Students apply for an aegrotat pass when they are unable to attend an assessment or their performance has been impaired due to illness or other unforeseen circumstances. (ii) A student had a missing term test or final exam score for some other reason. (iii) A student received a total score for the course equal to zero. These students did not attempt any assessment item.

[4] Krieg and Uyar (2003) have only 223 observations.

Also noteworthy in FIGURE 1 is that the distribution of test scores is constrained to lie between 0 and 100. Amongst other problems, this will cause the errors associated with a linear regression specification to be heteroscedastic. We address this problem in two ways. First, we use OLS but estimate the standard errors using the heteroscedastic-robust White procedure. OLS has the advantage of facilitating interpretation of the coefficient estimates. Accordingly, these are the results we report in our paper. However, we also estimated the key regressions using the more statistically appropriate fractional logit procedure. The results were virtually identical.[5]

TABLE 1 provides a statistical summary of the students represented in our study. Approximately 55 percent of the sample derived from Introductory Microeconomics classes. By construction, the data set consists of exactly half term-test and half final exam results. TABLE 1 also breaks down the essay and MC scores by term-test and final exam. Both components show higher scores on the final exam.

The variable *GPA* reports the student's weighted grade point average for all courses outside of ECON 104 (Introductory Microeconomics) and ECON 105 (Introductory Macroeconomics) in the same year that the student was enrolled in the respective economics class. For example, if a student was enrolled in ECON 104 in Semester 1 of 2005, *GPA* reports their weighted grade point average for all courses they took in calendar year 2005, excluding ECON 104 and 105. Grade points range from -1 (for a letter grade of E = fail) to 9 (for a letter grade of A+). The variable *COMPOSITE* is a weighted average of the essay and MC components, and is used later in the study when we estimate Walstad and Becker (1994)-type regressions.

[5] The fractional logit results are available upon request from the authors.

While not reported in TABLE 1, approximately 56 percent of the sample is male. A little less than half of the students in our sample are New Zealand natives or of European extraction. Approximately 43 percent of the students are Asian. This high percentage is due to a surge in Asian enrollments that occurred in the early 2000's in New Zealand universities. This tapered off substantially in the latter years of the sample. Maori, Pacific Islanders, and Others (primarily Africans and Middle Easterners) account for less than 8 percent of our sample. With respect to language, most of the sample declared English as their "first language." Even so, a little less than 40 percent declared that English was not their "first language," with the great majority of these identifying with Chinese.

III. RESULTS

The first step of our analysis consists of determining to what extent performance on the essay component of an assessment is "predictable" from the student's MC score on that assessment. If the corresponding regressions produce R^2 values close to one, this would clearly indicate that essay scores added little information to that already provided by the student's MC performance. We could then confidently conclude that essay and MC questions measured the same thing.

TABLE 2 summarizes the results of this analysis. We divided our data set into four, mutually exclusive sets of observations: (i) term tests and (ii) final exams from Introductory Microeconomics classes; and (iii) term tests and (iv) final exams from Introductory Macroeconomics classes. For each sample, we regressed students' essay scores on their MC scores for the same assessment. In addition, we aggregated all the observations into one sample. Not surprisingly, we find that MC scores are significant predictors of students' essay scores. An extra point on the MC component predicts an additional 0.7 to 1.1 points on the essay component, depending on the sample.

On the other hand, we also find that the R^2 values are never close to 1. Interestingly, MC scores are more successful at predicting essay scores for final exams: The R^2 values for the final exam regressions are close to 50 percent, while those for the term tests are in the low- to mid-30's.[6] For the full sample, the R^2 of the regression of essay scores on MC scores was a little less than 40 percent.[7]

To facilitate comparison with other studies, the last line of the table reports the simple correlation between essay and MC scores. Walstad and Becker (1994, page 194) report simple correlations of 0.69 and 0.64 for the Micro and Macro AP tests. Lumsden and Scott (1984, page 367) report correlations of 0.18 and 0.26 for introductory Micro and Macro courses, respectively. In contrast, they cite a number of other studies where the correlations range higher, though still lower than reported here. Thus, our finding that essay scores are far from being perfectly, or even near perfectly, predictable from MC scores appears to be the norm.

Unfortunately, while an R^2 close to 1 provides strong evidence that essay and MC questions measure the same thing, it is unclear what an R^2 far from 1 implies. Is the unexplained component in essay scores due to the fact that essay questions measure something different than MC questions? Or are the two question-types assessing the same thing(s) but with measurement error?

[6] Conventional wisdom is that essay questions are "noisier" assessments, a view supported by the fact that essay scores have greater dispersion (cf. FIGURE 1). One might suppose that the greater predictability of essay scores on final exams was due to students learning better essay-writing skills over the course of the semester (e.g., learning what the lecturer was looking for grading answers). However, this hypothesis is not supported by the data. While the standard deviation of the multiple-choice scores decreased from term-test to final exam (15.7 versus 14.7), they increased for essay scores (20.4 versus 21.3).

[7] We also investigated the effect of including higher-order, polynomial terms for the MC variable. This added little to the overall explanatory power of the equations.

If we had an alternative measure of student learning, we could take the residuals from the regressions in TABLE 2 and test if they were independent predictors of academic achievement. If the residuals were unrelated to student learning, say were pure measurement error, then one would expect them to be unrelated to this alternative measure. Alternatively, if we could show that these residuals were positively related to this alternative measure, this would provide evidence that the residuals contained independent "information about student achievement" that was not captured by MC responses.

Unfortunately, we do not have an alternative measure of student learning for the same assessment. We do, however, have a close substitute. Because we have repeated observations for each student, we can test whether residuals from the term test regressions are related to achievement on the final exam. If the residuals represent pure measurement error, one would not expect to find any relationship with students' final exam performance.

Column (1) of TABLE 3 reports the results of a regression where students' essay scores from the final exam were regressed on (i) their MC scores from the term test, and (ii) the unexplained component of their essay score from the term test (i.e., the residual from the regression specification that was reported in TABLE 2).[8] We separate the 2002-2006 and 2007 final exams because the 2007 final exams included a larger share of material from the first half of the course. We also separate the Introductory Microeconomics and Introductory Macroeconomics final exams. In each of the six

[8] The residual variables come from term-test essay regressions using the same observations as the TABLE 3 samples (e.g., "All Observations (2002-2006)," "All Observations (2007)," etc.)

samples, the *Residual* variable has very large *t*-values. In addition, the respective coefficients are all positively-signed.

This finding is consistent with the hypothesis that essay scores contain unique information about student learning. However, it is possible that the predictive component of term-test essay scores may be related to something other than learning outcomes. For example, suppose students with bad handwriting receive lower marks on essay tests, ceteris paribus. Then a lower score on the term-test essay could be predictive of a lower score on the final exam essay because it was predictive of bad handwriting.

To check this possibility, we also regressed students' final exam MC scores on the same two variables used to predict their final exam essay scores. The qualitative results remain unchanged. For each sample, the *Residual* variable is positively correlated and highly, statistically significant. In other words, the unexplained component of term-test essay scores predicts student achievement on both the (i) essay and (ii) MC components of the final exam.

The unique nature of our data set also allows us to undertake another test. Our data set includes information on students' grades in every course they have taken at the University of Canterbury. We use this information to calculate a GPA value based on their performance in non-introductory economics classes. For example, suppose a student took Introductory Microeconomics (ECON 104) in the first semester of 2005. We calculate their GPA over all other courses during the 2005 academic year, excluding their performance in ECON 104. If they subsequently took Introductory

Macroeconomics (ECON 105) in the second semester of 2005, we also exclude their performance in that class.[9]

It is well-known that student achievement is correlated across classes. This is consistent with the idea that student inputs into academic achievement, such as intelligence and good study habits, are transferrable across classes. If essay scores measure a component of learning that is not assessed by MC scores, then one might expect this independent information to be predictive of learning in other classes. Note that this need not be the case. It could be that essay scores in economics classes are only predictive of learning outcomes that are associated with economics content. But a positive finding relating the *Residual* variable with achievement in other courses would be further evidence that essay scores measure learning-related outcomes not captured by MC responses.

TABLE 4 reports the results of regressing students' *GPA* on their performance in a given introductory economics class. We employ four measures of student achievement: the student's MC score on the (i) term-test and (ii) final exam in that course; and the residuals from the (iii) term-test and (iv) final exam essay regressions, also from that course. These latter two variables are generated from TABLE 2-type regressions. They represent the component of the student's essay score that cannot be explained by their MC performance on the same assessment. We divide our observations into the same six samples that we used in TABLE 3.

[9] We chose to exclude both introductory economics classes because of similarities in the way the two classes were assessed. Since the two lecturers work closely together, it is possible that their assessment styles were similar. Correlation in performance across the two classes might represent students' ability to perform well on a particular style of assessment, and not an independent observation about student learning outcomes.

For each sample, we investigate whether the individual *Residual* variables are positively and significantly related to their outside *GPA* values. We also perform an *F*-test of the joint significance of the two *Residual* variables. Once again, the results in every case are consistent with the hypothesis that essay scores measure independent information not captured by students' MC scores. An extra "unexplained" point on the essay component of an assessment is associated with an increase in their outside *GPA* of anywhere from 0.0064 (cf. Column 3, Sample 3b) to 0.0662 points (cf. Column 4, Sample 3b). Interestingly, final exam performance seems to be a better predictor of outside *GPA* for both MC and essay scores. Furthermore, the individual *Residual* variables are each statistically significant at generally high *t*-values. The joint *F*-tests all have *p*-values that indicate significance at the 0.01% level.

Taken together, the results from TABLES 2 through 4 provide strong evidence that the essay and MC questions in our data do not measure the same things. Further, the latter two tables demonstrate that the component of essay scores that is not predictable from MC scores is positively and significantly related to academic achievement. While other studies, such as Kennedy and Walstad (1997), find evidence that essay and MC responses are "different," our study is the first to link these differences to learning outcomes.

IV. RELATING OUR FINDINGS TO THOSE OF PREVIOUS STUDIES

Our finding that essay and MC scores do not measure the same thing is at variance with a number of influential studies. In this section, we want to explore whether this is due to differences in our data, or differences in empirical procedures.

Bennett, Rock, and Wang (1991) and Thissen, Wainer, and Wang (1994) are widely-cited studies from the educational measurement literature. BRW base their analysis from a sample of responses from the College Board's Advancement Placement (AP) examination in Computer Science. TWW re-analyze BRW's data, and add a similar sample from the AP exam in Chemistry. Both employ common factor analysis to study the relationship between "free response" and MC questions.[10] Both find that a single factor explains most of the variation in the respective questions. They therefore conclude that these two question-types measure the same thing.[11]

While BRW and TWW employ factor analyses, they use somewhat different techniques. BRW use a model in which free response and MC questions are each loaded on a single factor. These two (correlated) factors are then analyzed to determine whether they contain unique information. In contrast, TWW employ a more general procedure to decompose the variation in the two types of questions into multiple factors.

The AP exam in Computer Science consists of 50 MC questions, and 5 free-response questions. The AP exam in Chemistry consists of 75 MC questions and four sections of free-response questions, some of which contain multiple problems. BRW and TWW break up the respective components into multiple "parcels." BRW re-organize the 50 MC questions into five sets "("parcels") of ten questions each. TWW convert the original 75 MC questions into fifteen, five-question parcels. These parcels become, in a sense, separate variables which are then decomposed into factors.

[10] For the difference between "free-response/constructed response" and "essay" questions, see footnote above.

[11] While both studies find more than one significant factor, they both conclude that a single factor is able to explain most of the variation in the two types of questions.

We attempt to replicate BRW's and TWW's factor analysis results. If we cannot replicate their results, this would suggest that our data are substantially different from theirs. In contrast, if are able to replicate their results, this would indicate that our different conclusions derive from different empirical procedures. Given the preceding evidence on essay and MC questions, it would suggest that the factor analysis approach is unreliable for determining whether essay and MC questions measure the same thing.

Unfortunately, our data contain fewer questions than BRW and TWW and are thus less amenable to "parcelization." Instead, we apply principal component analysis (PCA) to students' scores on the essay and MC components. PCA is related to factor analysis in that its "principal components" are akin to the factors identified by factor analysis. It has the advantage in that it produces a unique decomposition of the correlation matrix.[12] In contrast, factor analysis typically involves a subjective procedure ("rotation") that allows one to generate alternative sets of factors from the same data. A particularly attractive feature of PCA for our purposes is that it yields a straightforward measure of the amount of variation "explained" by each of the principal components.

TABLE 5 reports the results of applying PCA to the same five samples we previously analyzed in TABLES 2 and 3. As there are only two variables (Multiple-Choice and Essay), there are a total of two principal components. By construction, these two principal components explain all of the "variation" in the correlation matrix.

The first item of interest in TABLE 5 is the column of "eigenvalues." These provide a measure of importance for each of the principal components. In factor analysis, two common approaches for choosing the number of "factors" is Kaiser's eigenvalue rule

[12] Non-unique solutions can arise when two or more eigenvalues are exactly equal, but this is rarely encountered in practice.

and Cattell's scree test. The first of these selects factors having eigenvalues greater than one. The second of these plots the eigenvalues in decreasing order and selects all factors immediately preceding an abrupt leveling off of the values. Both approaches lead to the conclusion that there is one main factor underlying students' essay and MC responses in each of the samples. This finding is reinforced by the second column in TABLE 5. "Proportion" translates these eigenvalues into shares of total variation in the correlation matrix. These range from 78-85 percent across the different samples.

In summary, we find evidence (i) that a single factor underlies students' essay and MC responses in our data, and (ii) this single factor is able to explain most of the variation in the respective scores.[13] In other words, when we use an empirical procedure similar to what BRW and TWW employ, we are led to the same conclusion. This raises serious doubts about the appropriateness of factor analysis for addressing the question, "Do essay and MC questions measure the same thing?" Our analysis demonstrates that it is possible for this empirical procedure to produce a positive answer to this question, even when the underlying data contain strong, contrary evidence.

Walstad and Becker (1994) is another study that has been very influential in the debate over essay versus MC questions. Their study analyzes AP Microeconomics and Macroeconomics exams. Each of these has essay and MC components from which an overall composite score is formed, with the components receiving weights of two-thirds and one-third, respectively. WB use these data to regress the composite scores on the MC scores. They find that the MC scores explain between 90 and 95 percent of the variation in composite scores. WB conclude that there are "no differences, or only slight

[13] BRW conclude that one factor explains most of the variation by virtue of a battery of goodness-of-fit measures, finding that the second factor adds little in the way of goodness-of-fit. TWW reach this conclusion by noting that the factor loadings on the second factor are relatively small.

differences, in what the two types of tests and questions [multiple-choice and essay] measure."

Conveniently, WB report simple correlations between the essay and MC components of the AP exams. These fall in the same range as the correlations we report for our data in TABLE 2. Thus, it should not be surprising that we are able to produce WB-type regressions that are very similar to theirs.

We construct composite scores from the MC and essay components using the same weights as the AP exams. We then estimate WB-type regressions using the same five samples we used for our original analyses. TABLE 6 reports the results. Of interest here are the R^2 from the respective regressions. These range between 85 and 90 percent.[14] Using the same specification, WB obtained an R^2 of 94% for the Microeconomics exams, and an R^2 of 90% for the Macroeconomics exams. Our macro results are about the same as WB's, while our micro results are somewhat lower.

In conclusion, the strongest evidence that essay and MC questions measure the same thing comes from factor analysis and WB-style regressions. The preceding analysis argues that both these approaches are unreliable in the following sense: It is possible for these empirical procedures to produce an affirmative conclusion, even when the underlying data contain strong, contrary evidence.

In placing these studies in perspective, it is useful to recall the "policy question" that motivates them. If it could be shown that essay and MC questions measure the same thing, then instructors could get the same information about learning outcomes using an all-MC format, at lower total cost. Our analysis suggests that essay and MC questions do not measure the same thing. Yet, it could still be the case that an all-MC format is

[14] These results are very similar to those obtained by Krieg and Uyar (2001).

preferable if the extra information provided by essay questions was not sufficient to justify their higher costs.

This highlights two separate, but related research questions: (i) Do essay and MC questions measure the same thing?, and (ii) Are the benefits of essay questions sufficient to compensate their costs? Perhaps WB-style regressions are more appropriate for addressing this second question. If composite scores are near-perfectly predictable from MC scores, this may suggest that the benefits of essay questions are relatively small. However, even this conclusion does not necessarily follow. The slippage occurs in mapping R^2 values to benefits.

As Kennedy and Walstad (1997) point out, it is grades, not R^2 values, which matter to students and lecturers. KW use simulation exercises to estimate the effect of moving to an all-MC format for the AP test. They report that the number of students who would receive different AP grades is small but statistically significant. However, alternative simulation assumptions produce larger effects.

Like KW, we conclude that essay and MC questions do not measure the same thing. KW's approach has an advantage over ours in that they relate differences in essay and MC scores to an outcome that can be mapped into a benefit versus cost framework. The unique contribution of our study is that we provide evidence that these differences are related to student achievement.

V. CONCLUSION

Our study empirically investigates the relationship between essay and multiple-choice (MC) questions using a unique data set compiled from several years of university introductory economics classes. We find that MC questions are able to explain, at best, about 50 percent of the variation in essay scores. The main contribution of our study is that we show that the corresponding residuals are related to student learning. Specifically, we find that the component of essay scores that cannot be explained by MC responses is positively and significantly related to (i) performance on a subsequent exam in the same course, and (ii) academic performance in other courses.

A further contribution of our study is that we demonstrate that empirical approaches that rely on factor analysis or Walstad-Becker (1994)-type regressions are unreliable in the following sense: It is possible for these empirical procedures to lead to the conclusion that essay and MC questions measure the same thing, even when the underlying data contain strong, contrary evidence.

Further progress on the essay versus MC debate will likely come from more careful analyses of the benefits and costs of these two kinds of questions. Kennedy and Walstad (1997) show one way forward: Their paper models how one can compare grade distributions using alternative test formats. Another possible approach is to compare essay and MC scores on how well they predict future academic success. We hope this study stimulates future research efforts in this direction.

REFERENCES

Bennett, R., E., Rock, D., A., & Wang, M. (1991). Equivalence of Free-Response and Multiple-Choice Items. *Journal of Educational Measurement, 28(1),* 77-92.

Kennedy, P. E., & Walstad, W. B. (1997). Combining Multiple-Choice and Constructed Response Test Scores: An Economists View. *Applied Measurement in Education, 10(4),* 359-375.

Krieg, R., G., & Uyar, B. (2001). Student Performance in Business and Economic Statistics: Does Exam Structure Matter? *Journal of Economics and Finance, 25(2),* 229-241.

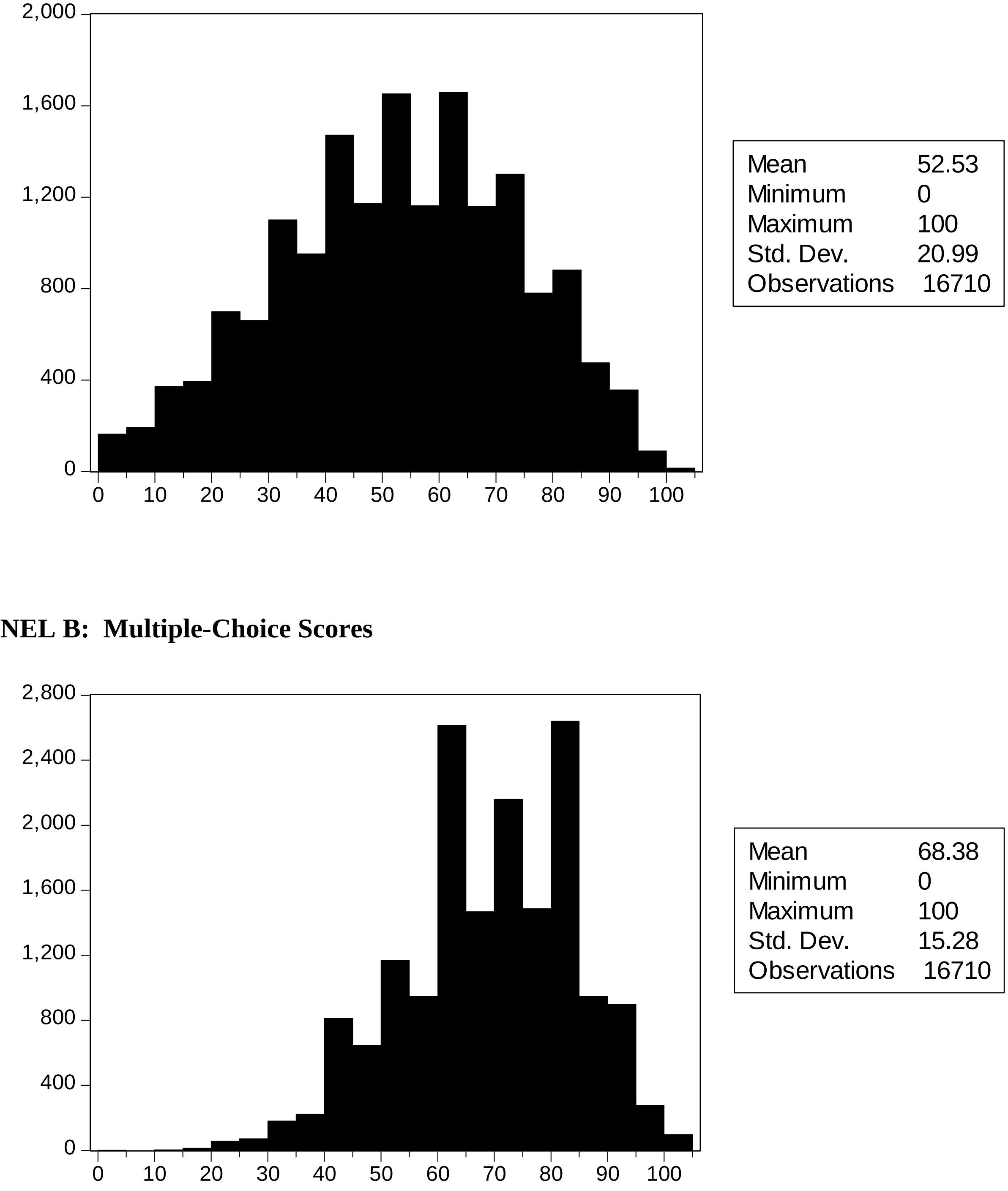
Lumsden, K.G, & Scott, A (1987). The Economics Student Reexamined: Male-Female Differences in Comprehension. *Journal of Economic Education, 18(4),* 365-375.

Thissen, D., Wainer, H., & Wang, X. (1994). Are Tests Comprising Both Multiple-Choice and Free-Response Items Necessarily Less Unidimensional Than Multiple-Choice Tests? An Analysis of Two Tests. *Journal of Educational Measurement, 31,* 113-123.

Wainer, H. & Thissen, D. (1993). Combining multiple-choice and constructed response test scores: Towards a Marxist theory of test construction. *Applied Measurement in Education, 6,* 103-118.

Walstad, W. B., & Becker, W. E. (1994). Achievement Differences on Multiple-Choice and Essay Tests in Economics. *American Economic Review, 84,* 193-196.

FIGURE 1
Statistical Summary of Multiple-Choice and Essay Scores



PANEL A: Essay Scores

PANEL B: Multiple-Choice Scores

TABLE 1
Statistical Summary of Data

VARIABLE	OBSERVATIONS	MEAN	MINIMUM	MAXIMUM	STD. DEV.
MICRO	16710	0.554	0	1	0.497
TERM_TEST	16710	0.500	0	1	0.500
ESSAY (Term Test)	8355	50.0	0	100	20.4
ESSAY (Final Exam)	8355	55.0	0	100	21.3
MULTIPLE-CHOICE (Term Test)	8355	66.8	0	100	15.7
MULTIPLE-CHOICE (Final Exam)	8355	69.9	16.7	100	14.7
GPA	16710	3.53	-1	9	2.49
COMPOSITE	16710	63.1	10	100	15.5

TABLE 2
Predicting Essay Scores Using Multiple-Choice Scores

			SAMPLE		
	Micro/Term Tests ***(1)***	***Micro/Final Exams*** ***(2)***	***Macro/Term Tests*** ***(3)***	***Macro/Final Exams*** ***(4)***	***All Observations*** ***(5)***
Constant	-7.4980 (-6.72)	-12.1581 (-11.69)	6.1509 (5.79)	-21.2494 (-18.03)	-6.0626 (-10.69)
Multiple-Choice	0.8097 (50.96)	0.9832 (67.81)	0.7143 (43.55)	1.0608 (67.28)	0.8568 (106.63)
Observations	4628	4628	3727	3727	16710
R^2	0.347	0.470	0.318	0.508	0.389
Simple Correlation	0.589	0.686	0.564	0.713	0.624

NOTE: Values in parentheses are *t*-statistics calculated using heteroscedastic-robust (White) standard errors.

TABLE 3
Predicting Final Exam Performance From Term Test Scores

VARIABLE	*Dep. Variable =* *Essay (Final Exam)* *(1)*	*Dep. Variable =* *Multiple-Choice (Final Exam)* *(2)*
<u>Sample (1a): ALL OBSERVATIONS (2002-2006)</u>		
Constant	7.5982 (9.55)	37.3361 (60.72)
Multiple-Choice (Term Test)	0.7152 (63.24)	0.4933 (57.12)
Residual from Term-Test Essay Regression	0.5292 (49.49)	0.3092 (38.97)
R^2	0.468	0.410
Observations	7270	7270
<u>Sample (1b): ALL OBSERVATIONS (2007)</u>		
Constant	-12.2469 (-5.97)	25.8495 (14.34)
Multiple-Choice (Term Test)	0.9591 (33.80)	0.6170 (25.09)
Residual from Term-Test Essay Regression	0.6331 (22.03)	0.2198 (11.80)
R^2	0.579	0.415
Observations	1085	1085
<u>Sample (2a): MICRO (2002-2006)</u>		
Constant	-0.6955 (-0.58)	27.7901 (30.76)
Multiple-Choice (Term Test)	0.7954 (48.93)	0.5879 (48.29)
Residual from Essay Regression	0.4710 (31.79)	0.2740 (25.20)
R^2	0.459	0.4424
Observations	3947	3947

VARIABLE	***Dep. Variable =*** ***Essay (Final Exam)*** ***(1)***	***Dep. Variable =*** ***Multiple-Choice (Final Exam)*** ***(2)***
Sample (2b): MICRO (2007)		
Constant	-12.7999 (-4.93)	23.3048 (11.25)
Multiple-Choice (Term Test)	0.9946 (26.90)	0.6108 (21.07)
Residual from Term-Test Essay Regression	0.6112 (17.21)	0.2547 (11.73)
R^2	0.578	0.454
Observations	681	681
Sample (3a): MACRO (2002-2006)		
Constant	9.6417 (8.77)	40.0442 (46.99)
Multiple-Choice (Term Test)	0.7335 (44.66)	0.5055 (39.99)
Residual from Term-Test Essay Regression	0.5757 (33.66)	0.2808 (22.50)
R^2	0.486	0.404
Observations	3323	3323
Sample (3b): MACRO (2007)		
Constant	-13.8856 (-4.07)	34.2929 (12.53)
Multiple-Choice (Term Test)	0.9375 (20.68)	0.5685 (15.96)
Residual from Term-Test Essay Regression	0.6663 (13.09)	0.3167 (9.80)
R^2	0.581	0.479
Observations	404	404

NOTE: Values in parentheses are *t*-statistics calculated using heteroscedastic-robust (White) standard errors.

TABLE 4
Predicting Student GPA Using Term Test and Final Exam Scores

	EXPLANATORY VARIABLES			
	Multiple-Choice (Term Test) (1)	***Multiple-Choice (Final Exam) (2)***	***Residual from Term-Test Essay Regression (3)***	***Residual from Final Exam Essay Regression (4)***
Sample (1a): ALL OBSERVATIONS (2002-2006)				
Estimated Coefficients	0.0317 (19.92)	0.0742 (41.73)	0.0269 (19.92)	0.0488 (34.76)
R^2 = 0.630 , Observations = 1085				
Hypothesis Test (Residuals = 0):	*F* = 1218.45 (*p*-value = 0.0000)			
Sample (1b): ALL OBSERVATIONS (2007)				
Estimated Coefficients	0.0301 (5.99)	0.0925 (19.63)	0.0188 (5.38)	0.0498 (13.68)
R^2 = 0.569 , Observations = 7270				
Hypothesis Test (Residuals = 0):	*F* = 195.92 (*p*-value = 0.0000)			
Sample (2a): MICRO (2002-2006)				
Estimated Coefficients	0.0356 (14.86)	0.0752 (30.93)	0.0209 (11.49)	0.0508 (26.92)
R^2 = 0.567 , Observations = 3947				
Hypothesis Test (Residuals = 0):	*F* = 589.44 (*p*-value = 0.0000)			

	EXPLANATORY VARIABLES			
	Multiple-Choice (Term Test) (1)	***Multiple-Choice (Final Exam) (2)***	***Residual from Term-Test Essay Regression (3)***	***Residual from Final Exam Essay Regression (4)***
Sample (2b): MICRO (2007)				
Estimated Coefficients	0.0301 (4.74)	0.0973 (15.50)	0.0215 (5.15)	0.0440 (9.14)
R^2 = 0.629 , Observations = 681				
Hypothesis Test (Residuals = 0):	*F* = 92.87 (*p*-value = 0.0000)			
Sample (3a): MACRO (2002-2006)				
Estimated Coefficients	0.0363 (15.39)	0.0705 (26.46)	0.0321 (15.64)	0.0464 (22.19)
R^2 = 0.577 , Observations = 3323				
Hypothesis Test (Residuals = 0):	*F* = 616.87 (*p*-value = 0.0000)			
Sample (3b): MACRO (2007)				
Estimated Coefficients	0.0327 (4.10)	0.0943 (11.31)	0.0064 (2.88)	0.0662 (10.84)
R^2 = 0.642 , Observations = 404				
Hypothesis Test (Residuals = 0):	*F* = 1218.45 (*p*-value = 0.0000)			

NOTE: Values in parentheses are *t*-statistics calculated using heteroscedastic-robust (White) standard errors.

TABLE 5
Summary of Principal Component Analyses

Sample (1): All Observations

Principal Component	*Eigenvalue*	*Proportion*
1	1.6236	0.812
2	0.3764	0.188

Sample (2): Micro/Term Tests

Principal Component	*Eigenvalue*	*Proportion*
1	1.5846	0.792
2	0.4154	0.208

Sample (3): Micro/Final Exams

Principal Component	*Eigenvalue*	*Proportion*
1	1.6855	0.843
2	0.3145	0.157

Sample (4): Macro/Term Tests

Principal Component	*Eigenvalue*	*Proportion*
1	1.5636	0.782
2	0.4364	0.218

Sample (5): Macro/Final Exams

Principal Component	*Eigenvalue*	*Proportion*
1	1.7129	0.856
2	0.2871	0.144

TABLE 6
Summary of Regressions Based on Walstad and Becker's (1994) Specification

	SAMPLE				
	Micro/Term Tests ***(1)***	***Micro/Final Exams*** ***(2)***	***Macro/Term Tests*** ***(3)***	***Macro/Final Exams*** ***(4)***	***All Observations*** ***(5)***
Constant	-2.4999 (-6.72)	-4.0527 (-11.69)	2.0503 (5.79)	-7.0831 (-18.03)	-2.0209 (-10.69)
Multiple-Choice	0.9366 (176.85)	0.9944 (205.76)	0.9048 (165.51)	1.0203 (194.11)	0.9522 (355.55)
Observations	4628	4628	3727	3727	16710
$\boldsymbol{R^2}$	0.862	0.891	0.871	0.896	0.876

NOTE: The dependent variable is a composite assessment score created by weighting the multiple-choice and essay components by 2/3 and 1/2, respectively. These are the weights used by the Advanced Placement Economics test that was analysed by Walstad and Becker (1994).